CN 11-2223/N　J Tsinghua Univ (Sci & Tech), 2009, Vol. 49, No. 3　http://qhxbw.chinajournal.net.cn

基于概率积分变换与似然比的高维相关性检验

刘晓曙

(厦门大学 金融系, 厦门 361005)

摘　要: 高维相关性的拟合优度检验是 Copula尤其是高维 Copula实践应用中遇到的复杂课题。为了解决这一难题, 该文将 Berkowitz在 2001年提出的基于概率积分变换的似然比检验单变量分布的方法推广用来做多变量分布的拟合优度检验, 特别是用来检验高维 copula的拟合优度。通过 Monte Carlo模拟对比实验, 结果表明基于概率积分变换的似然比检验统计量对高维相关性的拟合优度具有很强的检测能力, 同时对小样本数据同样有效。

关键词: 拟合优度; 概率积分变换; 似然比; Copula

中图分类号: O 1　　文献标识码: A

文章编号: 1000-0054(2009)03-0461-04

Multi-dimensional correlation test based on the probability integral translation and likelihood ratio

LIU Xiaoshu

(Department of Finance, Xiamen University, Xiamen 361005, China)

Abstract The goodness-of-fit test for multi-dimensional correlations is a complex problem. In order to solve this problem, this paper generalizes the likelihood ratio method based on the probability integral translation for a single stochastic variable distribution, which developed by Berkowitz in 2001, to the case of multi-variable distributions. Monte Carlo experiments show that the likelihood ratio test based on the probability integral translation is robust for the goodness-of-fit test for multi-dimensional correlations and is especially effective with small samples.

Key words goodness-of-fit; probability integral translation; likelihood ratio; Copula

在统计学领域, 多变量分布应用非常广泛。经典的多变量分布函数有多变量正态分布、双变量 gamma分布、双变量 Pareto分布等。但这些分布的最大缺点是要求边际分布属于同一族分布。

为解决这个问题, Copula提供了强有力的工具。Copula是用来描述多个随机变量间相依结构的统计方法。利用随机向量的边缘分布, Copula可以用来确定随机向量的联合分布。该理论方法由 Sklar于 1959年提出, 即 Sklar定理[1]:

对于多元分布函数 F, 如果它的边际分布函数 $F_1(x_1), F_2(x_2), \cdots, F_n(x_n)$均连续, 则其 Copula函数具有唯一形式:

$$F(x_1, x_2, x_3, \cdots, x_n) = C[F_1(x_1), F_2(x_2), \cdots, F_n(x_n)]. \tag{1}$$

由上述定理可以看到, 对于连续多变量分布函数, 变量间的相关结构可以通过 Copula函数表示出来, 因此可以将边际分布与联合分布分开来考虑, 并可灵活地选择边际分布的形式。Copula函数的另一个重要特点是: 对函数变量作非线性单调变换, 由 Copula函数所确定的一致性和相关性测度不会改变, 因而 Copula函数非常便于处理随机变量间的非线性相关问题。近年来, Copula在金融市间序列分析中取得了广泛的应用。到目前为止, 常见的 Copula主要有 3大类: 椭圆 Copula、Archimedean Copula与极值 Copula。

然而, Copula方法的不足在于缺乏一个学界普遍认同的优度拟合度量方法, 即如何判断选定的 Copula模型多大程度上刻画了数据间的相关性结构。信息标准, 例如 Aikaike信息标准(AIC)、似然函数值等, 常常在实证研究中用作模型选择方法。但是, 这类方法仅仅是次优的方法, 它们无法告诉人们选择的 Copula模型到底多大程度上拟合了数据, 同时, 这类方法也无法说明在拟合数据方面这个模型是否显著优于另一个模型; 因此, 需要对 Copula发展一套能弥补信息标准不足的拟合优度(GOF)检验方法。

对于单变量分布函数来说, GOF检验方法通

收稿日期: 2008-01-22

作者简介: 刘晓曙(1975—), 男(汉), 湖南, 博士研究生。

E-mail: liuxiaoshu@tsinghua.org.cn

常用著名的 Anderson-Darling 检验、Kolmogorov-Smirnov 检验、Cremer-von Mises 检验以及半定量的 Q-Q 图等。对于多变量分布函数来说, GOF 检验问题则要复杂得多。最近,学界提出了几种关于 Copula 的拟合优度检验方法。当然,这些方法都属于探索阶段,目前还没有就最优 Copula 选择方法达成普遍的、较一致的看法。1993年, Genest 和 Rivest 发展了一套基于经验选择最佳 Archimedean Copula 的方法[2]。后来 Diebold[3]、Breyman[4]分别基于 Rosenblatt[5]提出的概率积分变换给出检验统计量。所谓概率积分变换就是给定一个 d 维多变量分布函数,该变换将这些具有某种相关结构的随机变量变换为 $[0,1]^d$ 空间上相互独立的、均匀分布的随机变量。2006年 Genest 建立了基于 Kendall 过程的 GOF 检验方法[6]。同年, Panchenko 基于正定双线性形式提出了另一种 Copula 检验量[7]。以上方法中绝大多数方法都是将高维多变量问题通过降维方法转化为单变量问题,然后对单变量随机变量分布进行拟合优度检验。Berg 等发现这些检验方法存在或非一致性或有偏或数值计算无效率的问题,他们基于修正的概率积分变换提出另一种 Copula 拟合优度检验方法[8]。此外, Jadran 等发展了基于似然比的 copula 拟合优度检验方法[9]。

本文基于概率积分变换与似然比提出另外一种关于多变量相关性的拟合优度检验方法。与以上基于概率积分变换不同的是,该方法无需通过降维方法将高维多变量问题转化为单变量问题,而是直接对高维多变量问题直接处理。

1 基于概率积分变换与似然比的拟合优度检验方法

Berkowitz 针对单变量分布函数提出了基于概率积分变换与似然比的拟合优度检验方法[10]。本文在此将它推广到多变量分布函数情形。

1.1 条件概率积分变换

目前,学界已发展了几种概率积分变换。本文主要考虑 Rosenblatt 提出的变换[5],也称为条件概率积分变换(CPIT),即给定一个 d 维多变量分布函数,该变换将这些具有某种相关结构的随机变量变换为 $[0,1]^d$ 空间上相互独立的、均匀分布的随机变量。CPIT 普遍用来将任何一组来自已知随机变量分布函数的数据变换为一组独立的、均匀分布的数据。因此,如果有某种检验方法能有效地对多变量进行独立、均匀分布的拟合优度检验,那么, CPIT 就可以用作检验任何假设模型的的拟合优度。

条件概率积分变换定义:

令 $Z=(Z_1,Z_2,\cdots,Z_d)$ 是来自边际分布为 $F_i(z_i)=P(Z_i\leqslant z_i)$、条件分布为 $F_{i|1,\cdots,i-1}(Z_i\leqslant z_i|Z_1=z_1, Z_2=z_2,\cdots,Z_{i-1}=z_{i-1})$ 的随机变量矢量。其中 $i=1,\cdots,d$。Z 的 CPIT 定义为 $T(Z)=(T_1(Z_1),T_2(Z_2),\cdots,T_d(Z_d))$,其中:

$$
\begin{aligned}
&T_1(Z_1)=P(Z_1\leqslant z_1)=F_1(z_1),\\
&T_2(Z_2)=P(Z_2\leqslant z_2|Z_1=z_1)=F_{2|1}(z_2|z_1),\\
&\quad\vdots\\
&T_d(Z_d)=P(Z_d\leqslant z_d|Z_1=z_1,\\
&\qquad Z_2=z_2,\cdots,Z_{d-1}=z_{d-1}).
\end{aligned}
\tag{2}
$$

则随机变量 $V_i=T_i(Z_i)$, $i=1,2,\cdots,d$ 是 $[0,1]^d$ 空间上的独立均匀分布。

因此,可以给定一个多变量分布零假设,对原始数据应用 CPIT 变换,然后对变换后的数据进行 GOF 检验。此时, GOF 检验的本质就是如何检验数据在 $[0,1]^d$ 空间上是独立的均匀分布。基于 CPIT 的拟合优度检验的主要优点在于所有的零假设和备择假设都是一样的,而与 CPIT 变换前原始数据的分布函数无关。

1.2 似然比检验

对于单变量分布($d=1$)来说,检验经过 CPIT 变换后的数据具有独立同分布的性质有很多的方法,如基于观测密度函数与理论密度函数距离的 Kupier 检验、Kolmogorov-Smirnov 检验、Cremer-von Mises 检验等。但是,当 $d>1$ 时,情况变得非常复杂。本文将对 CPIT 变换后的数据运用似然比检验方法进行研究。

似然比检验由于目标函数的非连续性使得检验统计量不具有通常意义上的渐近性质,本文通过参数检验方法验证独立的 $U[0,1]^d$ 零假设比较困难。因此,本文对 CPIT 变换后的数据在做一个简单的正态分布,即 $\{X_i=\Phi^{-1}(V_i)\}$,这里 Φ^{-1} 是标准单变量逆正态变换,其中

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{y^2}{2}\right)\mathrm{d}y.$$

原因为: 1) 在计算上,该变换容易实现; 2) 容易计算 Gauss 似然函数以及构造似然比(LR)检验统计量; 3) 基于正态变换后的似然比检验具有非常好的统计性质以及有限样本行为。

由下面的引理可以知道,关于随机变量 Z 的分

布的零假设与随机变量 V 服从均值为 0 协方差为单位矩阵的正态分布函数的零假设具有等价性。

引理 1[10]: 假设 $f(Z)$ 是 Z 的真实密度函数，$\hat{f}(Z)$ 是 Z 的模型密度函数，$\mathrm{J}(V)$ 是 CPIT 变换后的随机变量的密度函数，$\mathrm{H}(V)$ 是均值为 0 协方差为单位矩阵的正态分布函数，则

$$\lg\frac{f(Z)}{\hat{f}(Z)} = \lg\frac{\mathrm{J}(V)}{\mathrm{H}(V)}. \tag{3}$$

从引理 1 可知，模型密度函数的精度完全保持到正态分布变换后数据。例如，对于某些点或者区域，如果 $f(Z) > \hat{f}(Z)$，则在均值为 0 协方差为单位矩阵的正态分布定义空间相应的点或者区间同样有 $\mathrm{J}(Z) > \mathrm{H}(Z)$，反之亦然。也就是说，可以通过对正态分布变换后的数据作假设检验而对原始数据得出结论。

此外，似然比检验中研究者可以灵活决定使用哪些、多少个限制条件。

1.3 基于概率积分变换与似然比的拟合优度检验方法

通过对原始数据作 CPIT 变换后，可以得到一组数据 $V_t = \{V_i\}$，再通过对变换后的数据作正态分布变换 $\{X_i = \mathrm{H}^{-1}(V_i)\}$，这里 H^{-1} 为标准单变量逆正态变换，其中

$$\mathrm{H}(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{y^2}{2}\right)\mathrm{d}y.$$

根据零假设，$V_t \sim$ i.i.d. $N(0, I_{d\times d})$，这里 $I_{d\times d}$ 为 d 维单位矩阵。对 V_t 的该假设的检验有很多种方法构建检验统计量。

例如，可以相对零假设，备择假设为模型是一阶自回归的、均值非零，协方差矩阵为非单位阵。数学表达形式为

$$V_t - u = A(V_{t-1} - u) + e_t. \tag{4}$$

从而零假设: $u = 0$, $A = O_{d\times d}$, $\mathrm{E}(e_t) = I_{d\times d}$。式中，$O$ 为零矩阵，符号 E 表示协方差。

显然，如果采取简单的去掉第 1 个观察数据的方法，则与式(3)相联系的似然函数是:

$$L(u, A, C) = -\frac{1}{2}d(N-1)\lg(2\pi) + \lg\det C + \sum_{t=2}^{N}(z_t - u - Az_{t-1} + Au)'C(z_t - u - Az_{t-1} + Au)_t. \tag{5}$$

其中: N 为观测数据的组数，C 为 e_t 的协方差矩阵。

为了联合检验零假设: $u = 0$, $A = O_{d\times d}$, $\mathrm{E}(e_t) = I_{d\times d}$，定义统计量，

$$\mathrm{LR} = -2(L(0, O, I_{d\times d}) - L(u, A, C)). \tag{6}$$

那么在零假设下，统计量 $\mathrm{LR} \sim \chi^2\left(\frac{3}{2}d(d+1)\right)$。

备择假设可以推广到任意 k 阶自回归滞后项情形，即

$$V_t - u = A_1(V_{t-1} - u) + A_2(V_{t-2} - u) + \cdots + A_k(V_{t-k} - u) + e_t. \tag{7}$$

这时，似然函数与似然比统计量的构造都与式(5)—(6)类似。

定义外积 $\vee$ 运算为: 若 $x = \begin{pmatrix} a \\ b \end{pmatrix}$，则 $x \vee x \equiv \begin{pmatrix} a^2 \\ ab \\ b^2 \end{pmatrix}$。其他高维矢量的 $\vee$ 运算定义类似。类似地，可以在备择假设中考虑非线性相关性，例如:

$$V_t - u = A_1(V_{t-1} - u) + \cdots + A_k(V_{t-k} - u) + \Gamma_1(V_{t-1} - u)\vee(V_{t-1} - u) + \cdots + \Gamma_k(V_{t-k} - u)\vee(V_{t-k} - u) + e_t. \tag{8}$$

如果样本数允许，可以考虑任意阶的非线性相关项。当似然比 LR 检验中包含越多的条件限制，就越接近非参数检验方法。原则上，非参数检验方法可以在大样本条件下拒绝一切对独立的高维正态分布偏离的假设。但是，非参数检验的缺点是对数据量的要求非常苛刻。似然比检验方法的优点之一就是可以根据样本数以及检验需要加以确定限制条件。

2 拟合优度检验能力验证

本节将评估基于概率积分变换的似然比统计量的检验能力。通过执行混合检验，可以了解到似然比统计量探测肥尾与偏峰性质的能力。若统计量检验能够区分 Gauss Copula 和 t-Copula，则显示统计量能检测到分布的肥尾特征；若统计量检验能够区分 Gauss Copula 和 Clayton-Copula，则显示统计量能检测到分布的偏峰特征。

混合检验的对象是将 Gauss Copula 和另一个 Copula 分布函数混合起来构成一个混合的 Copula:

$$C_{\mathrm{mix}} = (1-p)C_{\mathrm{g}} + pC_{\mathrm{a}}, \quad p \in [0, 1]. \tag{9}$$

其中: C_{g} 为 Gauss Copula; C_{a} 为备选分析的 Copula 函数，或 Clayton-Copula，或 t-Copula; p 为混合几率。如果 $p = 0$，则 C_{mix} 就是 Gauss Copula; 如果 $p = 1$，则 C_{mix} 就是 Clayton-Copula 或 t-Copula。若 $0 < p < 1$，以几率 $(1-p)$ 从 Gauss Copula 中抽

样,以 p 从备选 Copula中抽样。在金融环境里, C_{mix} 可以解释为资产组合在一些时间内的相关结构遵从 Gauss Copula, 而另一些时间内资产组合的相关结构遵从备选 Copula。

零假设: 混合 Copula为 Gauss Copula。

在此零假设下对 C_{mix} 作 CPIT与 Gauss正态分布变换,计算前边提出的基于概率积分变换的似然比检验统计量。为获得拒绝率和相应的检验能力曲线,本文重复 1 000次上面的过程, n 表示模拟的样本数目。Monte Carlo模拟实验结果见图 1与图 2。

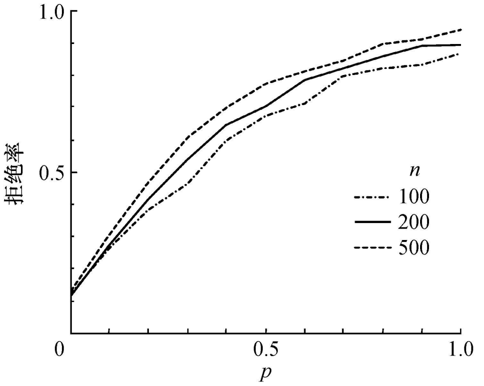

图 1 备择假设 C_{mix}为 Clayton-Copula

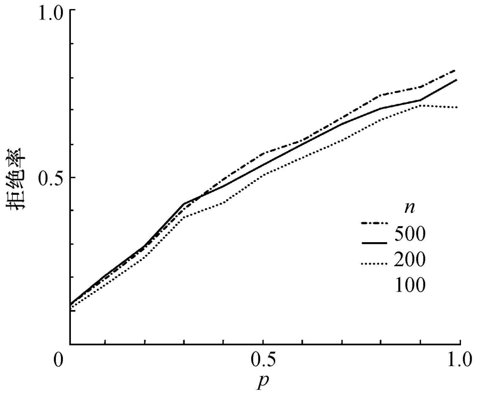

图 2 备择假设 C_{mix}是 t-Copula

从图 1与图 2可以看到,基于概率积分变换的似然比检验统计量具有很强的检测能力。而且,在小样本条件下依然具有很强的检测能力。

3 结论与讨论

本文将 Berkowitz提出的单变量分布基于概率积分变换似然比检验方法推广到高维分布的拟合优度检验方法中。通过 Monte Carlo模拟对比实验,发现基于概率积分变换的似然比检验统计量具有很强的 GOF检测能力。

此外,基于概率积分变换的似然比检验统计量还具有以下优点: 1) 容易计算; 2) 一般来说,直接对原始数据的分布作检验没有经济上的对应解释,而如果基于概率积分变换的似然比假设检验遭到拒绝,检验结果不会仅仅被动地给出拒绝信息,还可以根据假设遭到拒绝的原因提供建设性的指引。例如,如果假设由于相关性被拒绝,那么将出现持续的预测误差; 如果由于错误的平均假设,那么意味着过高(低)的平均预测。3) 基于概率积分变换的似然比检验统计量具有普适性,它与原始数据的分布没有直接联系,而且相对于其他基于矩的检验统计量来说,对小样本数据具有很强的检测能力。

当然,基于概率积分变换的似然比检验统计量不仅仅适用于高维相关结构函数的拟合优度的检验上,它还可以进一步对高维金融时间序列模型的设定正确性问题进行分析。

参考文献 (References)

[1] Sklar A. Fonctions de repartition a n dimensions et leurs marges [J]. *Publication de l' Institut de Statistique de l' Université de Paris*, 1959, **8**: 299 – 231.

Sklar A. Functions of repartition an dimensions marginal distributions [J]. *Journal of Statistics in Paris Univ*, 1959, **8**: 299 – 231. (in France)

[2] Genest C, Rivest L P. Statistical inference procedures for bivariate Archimedean copulas [J]. *Amer Statist Assoc*, 1993, **88**: 1034 – 1043.

[3] Diebold F X, Gunther T, Tay A S. Evaluating density forecasts with applications to financial risk management [J]. *International Economic Review*, 1998, **39**: 863 – 883.

[4] Breymann W, Dias A, Embrechts P. Dependence structures for multivariate high-frequency data in finance [J]. *Quantitative Finance*, 2003, **3**(1): 1 – 14.

[5] Rosenblatt M. Remarks on a multivariate transformation [J]. *The Annals of Mathematical Statistics*, 1952, **23**: 470 – 472.

[6] Genest C, Quessy J F, Remillard B. Goodness-of-fit procedures for copula models based on the probability integral transform [J]. *Scandinavian Journal of Statistics*, 2006, **33**: 337 – 366.

[7] Panchenko V. Goodness-of-fit test for copulas [J]. *Physics A*, 2005, **355**(1): 176 – 182.

[8] Daniel B, Henrik B. A copula goodness-of-fit approach based on the conditional probability integral transformation [EB/OL]. (2007-04-17). http://www.danielberg.no/publications/Btest.pdf.

[9] Dobri J, Schmid F. Testing goodness of fit for parametric families of copulas— Application to financial data [J]. *Simulation and Computation*, 2005, **34**: 1053 – 1068.

[10] Berkowitz J. Testing the accuracy of density forecasts in risk management [J]. *Journal of Business and Economic Statistics*, 2001, **19**: 465 – 474.